**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b7)**

**Measure Number** (*if previously endorsed*)**:** 0280

**Measure Title**: Dehydration Admission Rate (PQI 10)

**Date of Submission**: Click here to enter a date

**Type of Measure:**

|  |  |
| --- | --- |
| Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| Cost/resource | Process |
| Efficiency | Structure |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **2b7.** For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data),analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| abstracted from paper record | abstracted from paper record |
| administrative claims | administrative claims |
| clinical database/registry | clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| other: US Census Population Data | other: US Census Population Data |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

All analyses were completed using data from the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID), 2007-2012. HCUP is a family of health care databases and related software tools and products developed through a Federal-State-Industry partnership and sponsored by the Agency for Healthcare Research and Quality (AHRQ).1 HCUP databases bring together the data collection efforts of State data organizations, hospital associations, private data organizations, and the Federal government to create a national information resource of encounter-level health care data. The HCUP SID contain the universe of the inpatient discharge abstracts in participating States, translated into a uniform format to facilitate multi-State comparisons and analyses. All states provide data for community hospitals and together, the SID encompass about 97 percent of all U.S. community hospital discharges (in 2012, 46 states participated for a total of about 34 million hospital discharges from community hospitals). As defined by the American Hospital Association, community hospitals are all non-Federal, short-term, general or other specialty hospitals, excluding hospital units of institutions. Included among community hospitals are specialty hospitals such as obstetrics–gynecology, ear–nose–throat, orthopedic, pediatric institutions, short-stay rehabilitation, and long-term acute care.  Also included are public hospitals and academic medical centers. In the 2012 HCUP SID databases, 97.4% of all discharges are from community hospitals. Some states also include additional hospital types, which make up the remaining 2.6% of discharges, specifically psychiatric facility, alcohol and drug dependency facilities, and military hospitals.

The SID data elements include ICD-9-CM coded principal and secondary diagnoses and procedures, additional detailed clinical and service information based on revenue codes, admission and discharge status, patient demographics, expected payment source (Medicare, Medicaid, private insurance as well as the uninsured), total charges and length of stay ([www.hcup-us.ahrq.gov](http://www.hcup-us.ahrq.gov)).

**The area universe is defined as the county of the residence of the patient for discharges in the hospital universe.** The hospital universe is defined as all hospitals located in the U.S. that are open during any part of the calendar year and included in the SID database (see description above).

As noted, 97.4% of discharges in the 2012 SID are from “community hospitals.” The AHA defines community hospitals as follows: "All non-Federal, short-term, general, and other specialty hospitals, excluding hospital units of institutions." Starting in 2005, the AHA included long-term acute care facilities in the definition of community hospitals. These facilities provide acute care services to patients who need long-term hospitalization (stays of more than 25 days, but with an average stay of less than 30 days).

For the purpose of these analyses visits made by individuals residing in states that are not included in the HCUP databases for excluded from numerator counts.

Population estimates are derived from the US Census and are detailed in the 2013 Population File for use with the AHRQ Quality Indicators posted on the AHRQ QI website: <http://www.qualityindicators.ahrq.gov/Downloads/Software/SAS/V45/AHRQ%20QI%20Population%20File%20V4.5.pdf> and provided in the supplemental materials. Public-use files of intercensal and postcensal estimates of county-level population by five-year age group, sex, race, and Hispanic origin were acquired from the Census Bureau (<http://www.census.gov/popest/>) covering the years 1995 through 2011.

1HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2007-2012. Agency for Healthcare Research and Quality, Rockville, MD. [www.hcup-us.ahrq.gov/sidoverview.jsp](http://www.hcup-us.ahrq.gov/sidoverview.jsp). (AHRQ QI Software Version 4.5)

**1.3. What are the dates of the data used in testing**? 2007-2012

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| individual clinician | individual clinician |
| group/practice | group/practice |
| hospital/facility/agency | hospital/facility/agency |
| health plan | health plan |
| other: Geographic area, county | other: Geographic area, county |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

**Table 1. Reference Population Rate and Distribution of Area Performance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Overall Reference Population Rate** | | | | |
| **Year** | **Number Counties (Areas)** | **Outcome of Interest**  **(Numerator)1** | **Population at Risk**  **(Denominator)1** | **Observed Rate**  **Per 100,0001** |
| 2012 | 2,935 | 255,023 | 230,988,781 | 110.40 |
| 2011 | 3,018 | 276,293 | 230,827,273 | 119.70 |
| 2010 | 3,015 | 278,693 | 228,371,155 | 122.04 |
| 2009 | 2,863 | 298,685 | 223,703,795 | 133.52 |
| 2008 | 2,774 | 341,790 | 219,039,613 | 156.04 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Distribution of Area-level Observed Rates in Reference Population** | | | | | | | | |
| Year | Number of  Counties | Distribution of Observed Area-level Rates per 100,000 population (p=percentile)2 | | | | | | |
| Mean | SD | p5 | p25 | Median | p75 | p95 |
| 2011 | 2,862 | 137.56 | 69.58 | 33.29 | 88.44 | 129.95 | 180.27 | 268.46 |
| 2012 | 2,795 | 120.93 | 61.82 | 26.65 | 77.76 | 115 | 158.73 | 236.46 |

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2008-2012. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

1The observed rate refers to the total rate for all observations included in the reference population data (numerator) divided the total combined population of all counties included in the reference population data (denominator).

2The distribution of area rates reports the mean and standard deviation (SD) of the observed rates for all counties included in the dataset, as well as the observed rate for counties in the 5th, 25th, 50th (median), 75th, and 95th percentile. Note: Counties with rates outside of 1.5\*interquartile range are excluded as outliers.

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

See 1.5 (Table 1)

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

Table 1 includes 2012 data. All other testing use 2007-2011 data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)  
 **Performance measure score** (e.g., *signal-to-noise analysis*)  
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

We evaluate the reliability of risk-adjusted rates using a signal-to-noise ratio metric. The unit of analysis is the county. The noise measure is defined as the within-county variance, which is computed using a formula for the approximation for the variance of adjusted rates that is based on the observed and expected rates (See empirical methods document for details, or, for example Hosmer and Lemeshow, Stat Med. 1995). The signal measure is defined as the variance between counties. Signal-to-noise is computed for each county as (signal / (signal + noise)). The overall signal-to-noise ratio is the weighted average of the county-specific signal-to-noise ratios where the weights are defined as (1/(signal+noise)^2).

[Hosmer D and Lemeshow S. Confidence Interval Estimates of an Index of an Index of Quality Performance Based on Logistic Regression Models. Statistics in Medicine, Volume 14, 2161-2172 (1995).]

**2a2.3. For each level of testing checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Size Decile** | **Number**  **of Counties (Areas)** | **Ave. Number of**  **Persons per County (Area)**  **in Decile** | **Ave. Signal-to-Noise**  **Ratio for Counties (Areas)**  **in Decile** | **Percent of Signal**  **Variance Explained**  **by Performance Score** |
| 1 | 312 | 2,278.9 | 0.68591 | 0.83966 |
| 2 | 311 | 5,658.6 | 0.85517 | 0.89751 |
| 3 | 311 | 8,818.1 | 0.90016 | 0.92233 |
| 4 | 311 | 12,641.6 | 0.92673 | 0.93938 |
| 5 | 311 | 17,290.0 | 0.94452 | 0.95207 |
| 6 | 312 | 23,990.9 | 0.95852 | 0.96291 |
| 7 | 311 | 33,769.2 | 0.96954 | 0.97198 |
| 8 | 311 | 53,202.2 | 0.97994 | 0.98104 |
| 9 | 311 | 103,763.5 | 0.98887 | 0.98923 |
| 10 | 311 | 500,107.3 | 0.99630 | 0.99635 |
| Overall | 3,112 | 76,111.5 | 0.97886 | 0.98927 |

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

We evaluated the signal-to-noise ratio across deciles of county population size. The overall reliability of the risk adjusted measure is high. More than 80% of counties had signal-to-noise ratios that exceeded a threshold reliability of 0.80. Reliability was lower in counties with smaller populations. The observed signal-to-noise ratios were below 0.80 in counties with populations less than about 2,300 persons. This finding implies that when rates are smoothed, the rates for counties with smaller populations will be shifted more toward the overall (national) population rate.

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Overall the risk-adjusted rate is strongly reliable. Based on a norm of a signal-to-noise ratio of 0.80, 80% of areas exceed the norm. Reliability is less than the norm in areas with population less than approximately 2,300 persons, meaning that the performance score is reliability adjusted closer to the shrinkage target in those areas.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

**Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

**Temporal trends of the complexity of numerator cases**

We examined the complexity of the numerator cases from 2008 – 2012 using three metrics: 1) the percent of discharges with comorbidities can increase the risk of dehydration, 2) the mean number of comorbidities as defined by the AHRQ Comorbidity Index and 3) the mean age in years. The purpose of this analysis is to determine whether as rates of dehydration hospitalizations have decreased over time, the complexity of the remaining numerator cases has increased over time. Each of the three complexity metrics was calculated for each county in the SID; the distribution of the metrics across all counties is provided in Table 5.

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)

**Table 5. Temporal trends in numerator complexity (PQI 10)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2008 | | 2009 | | 2010 | | 2011 | | 2012 | |
| Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| Percent of discharges with comorbidity1 | | | | | | | | | | |
| Heart Failure | 11.8% | 8.9% | 11.7% | 9.6% | 11.8% | 10.0% | 12.2% | 10.9% | 12.1% | 10.4% |
| Diabetes | 26.1% | 12.7% | 26.5% | 13.6% | 27.1% | 13.6% | 28.3% | 14.7% | 28.4% | 14.8% |
| Renal Failure | 3.2% | 5.9% | 3.3% | 6.3% | 3.4% | 5.6% | 2.9% | 5.3% | 2.8% | 5.4% |
| Cancer | 13.4% | 10.7% | 13.7% | 11.1% | 13.7% | 11.9% | 13.8% | 11.4% | 13.9% | 13.0% |
| Mean number comorbidities1 | | | | | | | | | | |
| Medical comorbidities | 1.63 | 6.1 | 1.69 | 6.5 | 1.72 | 6.1 | 1.78 | 6.3 | 1.81 | 6.3 |
| Behavioral health comorbidities | 0.21 | 0.42 | 0.22 | 0.44 | 0.23 | 0.43 | 0.24 | 0.46 | 0.25 | 0.48 |
| Mean age in years | | | | | | | | | | |
| Mean age in years | 67.3 | 0.13 | 66.8 | 0.14 | 66.7 | 0.16 | 67.1 | 0.16 | 66.9 | 0.17 |

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2008-2012.

1AHRQ Comorbidity Index, version 3.7. Available at <http://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp> Medical comorbidities included: heart failure, valvular disease, pulmonary circulation disease, hypertension, paralysis, other neurological disorders, chronic pulmonary disease, diabetes without complications, diabetes with chronic complications, hypothyroidism, renal failure, liver, chronic peptic ulcer disease, HIV/AIDS, lymphoma, metastatic cancer, solid tumor, rheumatoid arthritis/collagen vascular diseases, coagulation deficiency. Behavioral health comorbidities included: alcohol abuse, drug abuse, psychoses, depression.

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)

From 2008-2012, there was less than 10% change in the percent of cases with all comorbidities known to have higher risk for dehydration except renal failure which demonstrated a 10% decrease. However, the mean number of comorbidities across counties increased by 11%. The number of numerator observations with a behavioral health comorbidity increased by 22% but the total number remained low (mean of 0.21 in 2008 compared to 0.25 in 2012). Mean age remained unchanged. From this, we conclude that there is only modest evidence of increasing complexity in the numerator population during the period of decrease, suggesting that the decrease does not simply reflect the shifting of non-complex patients from hospital care to other settings.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

**NA**  **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)  
 Not applicable

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)  
Not applicable

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)  
Not applicable

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

**Statistical risk model with** 27 **risk factors**

**Stratification by** Click here to enter number of categories **risk categories**

**Other,** Click here to enter description

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.   
Not applicable

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

For the area level indicators, the models use the following covariates: gender (male, female), age in 5-year age groups and an interaction with gender \* age. This is the default model and all risk adjusted data presented here are based on the default model. An optional risk adjustment model is available in the software, which also accounts for the percent of the population under the federal poverty level (data not shown).

**2b4.4. What were the statistical results of the analyses used to select risk factors?**The process to select risk factors is described in the AHRQ QI Empirical Methods report. The results of the analyses are provided in the PQI Parameter Estimates document. Both documents are found on the AHRQ QI website (www.qualityindicators.ahrq.gov) and provided in the excel spreadsheet provided with submission. The parameter estimates are derived from the reference population.

For the PQIs, risk adjustment is based on the age and gender mix of the population in the measured county. Age and gender is termed the “default risk adjustment model” in this documentation to distinguish this model from the alternative model which also accounts for poverty.

Risk Adjustment parameters and coefficients as presented in Table 6. The coefficients are also available in the *PQI Parameter Estimates* document (pages 12-13) posted at: <http://www.qualityindicators.ahrq.gov/Downloads/Modules/PQI/V45/Parameter_Estimates_PQI_45.pdf>.

**Table 6. Risk Adjustment Coefficients for PQI #10 Dehydration Admission Rate**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PARAMETER** | **LABEL** | **DF** | **ESTIMATE** | **STANDARD ERROR** | **WALD CHI-SQUARE** | **PR > CHI-SQUARE** |
| INTERCEPT |  | 1 | -4.6880 | 0.0079 | 347064.2 | < 0.0001 |
| SEX | Female | 1 | 0.0309 | 0.0096 | 10.24 | 0.0014 |
| AGE5 | Male, Age 18-24 | 1 | -3.6190 | 0.0183 | 39157.76 | < 0.0001 |
| AGE6 | Male, Age 25-29 | 1 | -3.4880 | 0.0202 | 29699.20 | < 0.0001 |
| AGE7 | Male, Age 30-34 | 1 | -3.4080 | 0.0201 | 28870.56 | < 0.0001 |
| AGE8 | Male, Age 35-39 | 1 | -3.2620 | 0.0189 | 29728.11 | < 0.0001 |
| AGE9 | Male, Age 40-44 | 1 | -3.0650 | 0.0172 | 31775.15 | < 0.0001 |
| AGE10 | Male, Age 45-49 | 1 | -2.7740 | 0.0150 | 34199.01 | < 0.0001 |
| AGE11 | Male, Age 50-54 | 1 | -2.4910 | 0.0137 | 33059.47 | < 0.0001 |
| AGE12 | Male, Age 55-59 | 1 | -2.2310 | 0.0132 | 28727.89 | < 0.0001 |
| AGE13 | Male, Age 60-64 | 1 | -1.9870 | 0.0128 | 23989.69 | < 0.0001 |
| AGE14 | Male, Age 65-69 | 1 | -1.6200 | 0.0127 | 16333.08 | < 0.0001 |
| AGE15 | Male, Age 70-74 | 1 | -1.2800 | 0.0126 | 10300.12 | < 0.0001 |
| AGE16 | Male, Age 75-79 | 1 | -0.9050 | 0.0123 | 5404.61 | < 0.0001 |
| AGE17 | Male, Age 80-84 | 1 | -0.5240 | 0.0121 | 1869.06 | < 0.0001 |
| AGE5 | Female, Age 18-24 | 1 | 0.1514 | 0.0245 | 38.15 | < 0.0001 |
| AGE6 | Female, Age 25-29 | 1 | 0.1473 | 0.0271 | 29.48 | < 0.0001 |
| AGE7 | Female, Age 30-34 | 1 | 0.2257 | 0.0264 | 73.21 | < 0.0001 |
| AGE8 | Female, Age 35-39 | 1 | 0.2560 | 0.0246 | 107.91 | < 0.0001 |
| AGE9 | Female, Age 40-44 | 1 | 0.2575 | 0.0223 | 133.15 | < 0.0001 |
| AGE10 | Female, Age 45-49 | 1 | 0.2162 | 0.0194 | 123.61 | < 0.0001 |
| AGE11 | Female, Age 50-54 | 1 | 0.1582 | 0.0178 | 78.94 | < 0.0001 |
| AGE12 | Female, Age 55-59 | 1 | 0.1354 | 0.0170 | 63.05 | < 0.0001 |
| AGE13 | Female, Age 60-64 | 1 | 0.1525 | 0.0165 | 85.38 | < 0.0001 |
| AGE14 | Female, Age 65-69 | 1 | 0.1201 | 0.0163 | 54.39 | < 0.0001 |
| AGE15 | Female, Age 70-74 | 1 | 0.1311 | 0.0161 | 66.62 | < 0.0001 |
| AGE16 | Female, Age 75-79 | 1 | 0.1137 | 0.0155 | 53.72 | < 0.0001 |
| AGE17 | Female, Age 80-84 | 1 | 0.0944 | 0.0150 | 39.39 | < 0.0001 |

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)

This analysis evaluates how strongly the risk adjustment model is associated with the event of interest (i.e. hospitalization for dehydration). The measure of discrimination, how well the risk adjustment model distinguishes events from non-events, is the c-statistic. The c-statistic is computed by assigning each observation a predicted probability of the outcome from the risk-adjustment model based on the value of the observations covariates (age and sex). Two copies of the dataset are sorted, first from highest to lowest predicted probability and second from lowest to highest predicted probability. This creates a set of pairs of observations. Pairs that consist of one event and one non-event (discordant pairs) are kept and concordant pairs are discarded. The c-statistic is a measure of the proportion of discordant pairs of observations for which the observation with the event had a higher predicted probability from the risk-adjustment model than the non-event. We did not employ common “goodness of fit” tests because these tests tend to not be informative with large samples.

We also evaluated the calibration of the risk adjustment model by evaluating how closely observed and predicted rates compare across deciles of the predicted rate. This analysis splits the sample into deciles based on predicted rates, and then compares these rates with the observed rates for the population in each decile. A well calibrated model, or one that does not over or under-estimate risk, will have comparable observed and predicted rates across the risk spectrum.

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***If stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**See 2b4.8 (Table 7)

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):   
See 2b4.8 (Table 7)

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

Table 7. Model Discrimination and Calibration

|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted**  **Rate Decile** | **Number of Persons**  **per Decile** | **Predicted**  **Rate** | **Observed**  **Rate** |
| 1 | 23,738,689 | 0.000345 | 0.000358 |
| 2 | 23,738,625 | 0.000403 | 0.000429 |
| 3 | 23,848,478 | 0.000491 | 0.000504 |
| 4 | 23,630,107 | 0.000559 | 0.000606 |
| 5 | 23,739,077 | 0.000567 | 0.000597 |
| 6 | 23,741,848 | 0.000625 | 0.000651 |
| 7 | 23,734,511 | 0.000707 | 0.000757 |
| 8 | 23,749,085 | 0.000752 | 0.000789 |
| 9 | 23,735,780 | 0.000791 | 0.000794 |
| 10 | 23,730,269 | 0.000955 | 0.000987 |
| C-statistic | | 0.709 | |

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

**2b4.9. Results of Risk Stratification Analysis**:

Not applicable

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

The model has moderate discrimination based on a norm c-statistic of 0.71 and moderate calibration.

**2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Not applicable

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

An indicator relies on historical data to predict future performance of a county. This analysis assesses the stability of rates over time and assesses the probability that an area is higher or lower than a benchmark or threshold. The analysis is reported by decile demonstrating performance across counties of various sizes. Each county is assumed to have an underlying distribution of smoothed rates that follows a Gamma distribution. The parameters of a Gamma distribution are shape and scale. For each county the shape is calculated as ((smoothed rate)2/ smoothed rate variance), and the scale is calculated as (smoothed rate variance / smoothed rate). The smoothed rate variance (aka posterior variance) is calculated as the signal variance – (reliability weight \* signal variance). The reliability weight is calculated as (signal variance / (signal variance + noise variance)). Counties are ranked by size and grouped into 10 equal categories of size (deciles). Two rates from the reference population at the 20th and 80th percentile (termed Benchmark and Threshold) are compared to the Gamma distribution of the smoothed rates for each county to determine if the county rate is better or worse than the Benchmark and Threshold rates with 95% probability. We report the proportion of counties in each decile above and below the Benchmark and Threshold rates.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Table 8. Performance Categories by County (Area) Size Decile

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | |  | | | |
|  |  |  | **Benchmark** | | | **Threshold** | |
| **Size Decile** | **Number**  **of Counties (Areas)** | **Ave. # of**  **persons per County (Area)**  **in Decile** | **Proportion**  **Better** | | **Proportion**  **Worse** | **Proportion**  **Better** | **Proportion**  **Worse** |
| 1 | 312 | 2,278.9 | 0.02885 | | 0.51923 | 0.56731 | 0.16346 |
| 2 | 311 | 5,658.6 | 0.07074 | | 0.64630 | 0.52090 | 0.20900 |
| 3 | 311 | 8,818.1 | 0.08039 | | 0.66881 | 0.56270 | 0.24116 |
| 4 | 311 | 12,641.6 | 0.07396 | | 0.71383 | 0.58521 | 0.21543 |
| 5 | 311 | 17,290.0 | 0.08360 | | 0.73955 | 0.58842 | 0.22830 |
| 6 | 312 | 23,990.9 | 0.06090 | | 0.76282 | 0.60256 | 0.23397 |
| 7 | 311 | 33,769.2 | 0.07717 | | 0.74277 | 0.72347 | 0.14469 |
| 8 | 311 | 53,202.2 | 0.08682 | | 0.79100 | 0.70096 | 0.16720 |
| 9 | 311 | 103,763.5 | 0.09646 | | 0.78135 | 0.80064 | 0.12540 |
| 10 | 311 | 500,107.3 | 0.08682 | | 0.81994 | 0.91961 | 0.05466 |
| All | 3,112 | 76,111.5 | 0.07455 | | 0.71851 | 0.65713 | 0.17834 |
| Patient weighted |  |  | 0.07315 | | 0.81809 | 0.86298 | 0.08164 |

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 4.5)

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

When examining the performance of the measure across deciles of county population size, we observed that counties with large populations were not more likely to be identified as better than the Benchmark compared to counties with small populations. In contrast, counties with large populations were less likely to fall below the Threshold than counties with smaller populations.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
Not applicable

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)  
Not applicable

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)  
Not applicable

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS**

**2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps―do not just name a method; what statistical analysis was used*)  
Not applicable

**2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data?** (*e.g.,**results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)  
Not applicable

**2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased** due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Not applicable